DOCUMENT RESUME

ED 127 337 TM 005 422

AUTHOR Rosmann, Michael R.
TITLE Statistical Analyses of Repeated Measures Designs: Problems and Some Solutions.
PUB DATE [73]
NOTE 58p.; Revision of a paper presented at the Annual Meeting of Southeastern Psychological Association (21st, Atlanta, Georgia, March 27-29, 1975)

EDRS PRICE MF-$0.83 HC-$3.50 Plus Postage.
DESCRIPTORS *Analysis of Variance; Correlation; Goodness of Fit; Hypothesis Testing; Mathematical Models; Measurement Techniques; Nonparametric Statistics; Observation; Research Design; *Statistical Analysis; Tests of Significance; Transformations (Mathematics); *Trend Analysis
IDENTIFIERS *Repeated Measures; Time Series Analysis

ABSTRACT

When repeated measures are obtained on the same subjects, interobservation dependencies frequently are generated. The major ways in which these dependencies can arise are illustrated and it is shown how these dependencies may invalidate the use of analysis of variance (ANOVA) and its extension, trend analysis, as methods of evaluating the data of such experiments. Several alternative procedures which involve ANOVA and trend analysis and which may avert interobservation dependencies are reviewed. The use of time series analysis with repeated measures experiments is also examined. Finally the advantages and limitations of graphic display only of data and directional correlation are discussed briefly. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED127337

Statistical Analyses of Repeated Measures Designs: Problems and Some Solutions

Michael R. Rosmann

University of Virginia

Running head: Analyses of Repeated Measures

TM005 422

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Abstract

When repeated measures are obtained on the same subjects, interobservation dependencies frequently are generated. The major ways in which these dependencies can arise are illustrated and it is shown how these dependencies may invalidate the use of analysis of variance (ANOVA) and its extension, trend analysis, as methods of evaluating the data of such experiments. Several alternative procedures which involve ANOVA and trend analysis and which may avert interobservation dependencies are reviewed. The use of time series analysis with repeated measures experiments is also examined. Finally, the advantages and limitations of graphic display only of data and directional correlation are discussed briefly.

Statistical Analyses of Repeated Measures Designs:
Problems and Some Solutions

Researchers in the social and natural sciences and many other areas of scientific endeavor frequently use experimental designs in which repeated measures are made of each experimental unit (subject). Two particularly attractive features of repeated measures designs are that they allow for economy of subjects by exposing the same subjects to multiple treatments and they allow for the study of change over time by taking multiple observations of the same subjects within a time sequence. When experimental designs calling for repeated measures of the same subjects are employed, the observations frequently are dependent (i.e. errors are correlated). This consequence jeopardizes the application of most statistical techniques readily available to researchers in the social sciences and other scientific areas because much of the statistical methodology developed thus far is concerned with models in which the observations are assumed to vary independently. In fact, nearly every model of analysis of variance (ANOVA) presented in such well known statistics texts as those authored by Ferguson (1971), Hays (1963) and McNemar (1969) and most other tests for differences between groups of scores (e.g., "student's t" Kurtz, 1965) assume that the observations are independent from one another.

Namboodiri (1972) reviewed several repeated measures designs which were constructed so as to either minimize cor-

relation among repeated observations or enable statisticians to partial out undesirable dependency effects. A fair appraisal of Namboodiri's review indicates that he incompletely evaluated sources of interobservation dependency and the possible means of "handling" these dependencies statistically. Furthermore, he provided few hints about evaluating data in which dependencies among the observations are an important desirable aspect rather than an annoying detraction.

This paper illustrates several common but often overlooked ways in which repeated observations of the same subjects can be correlated. By summarizing and extending the works of several statisticians, it is shown how these interobservation dependencies can invalidate the application of parametric ANOVA and its extension, trend analysis. Seven alternative statistical procedures involving the ANOVA and six alternatives involving trend analysis are presented. Time series analysis is also presented as an alternative statistic for evaluating the data of many experiments entailing repeated measures of the same subjects.

We begin this elucidation of several of the ways in which repeated observations of the same subjects can be correlated by illustrating the following longitudinal experiment. Suppose that an investigator wants to measure the frequency of verbal communication between married spouses in a nondisrupted situation when the partners can communicate freely and in a disrupted situation when interspouse communi-

cation is partially restricted. The investigator places the spouses in separate rooms and instructs them to communicate via button-operated intercoms which are hooked up to event recorders so as to tabulate the frequency of attempted communications by each person. The investigator allows the frequency of attempted verbal communications by the husband and wife to stabilize during a nondisrupted baseline condition. When the frequency of attempted communications by both persons has stabilized, the investigator then disrupts communication during an experimental condition so that only two out of every three messages sent by either spouse is actually received by the other spouse. The investigator disrupts the message-sending process so that the subjects remain unaware of the experimental manipulation. After a suitable experimental period the investigator reinstates baseline conditions. The hypothesis of this experiment is that the spouses will attempt a significantly greater number of communications during the disrupted condition than during the nondisrupted periods.

Following the experimental procedures above, suppose that the data shown in Figure 1 were obtained from a single

---

Insert Figure 1 about here

---

married couple during ten sessions conducted on ten consecutive days so as to minimize subject fatique. For the sake of maintaining simplicity of the statistical analyses, assume

also that there are 40 five-minute intervals per experimental condition. The data of this experiment, then, fit a 3 x 2 factorial matrix with an equal number (n) of repeated measures in each cell, such as is shown in Table 1.

Insert Table 1 about here

The data of this hypothetical longitudinal experiment are what might be expected during an A B A reversal of treatment conditions. They suggest relatively stable rates of attempted communications by both the husband and wife during nondisrupted Baseline conditions (Baseline 1 and 2). However, the rates of attempted communications accelerated highly when the husband and wife initially experienced the disrupted experimental condition (Experimental 1), then both rates dropped sharply and reached asymptote at rates below the baselines during the latter part of the experimental period.

In the example given, dependency among the observations occurs in the following ways:

1. The effect of a treatment appears to carry over into the period following the one in which the treatment was applied. Carryover effects appear to operate at two different instances. First, response bursts by both the husband and wife occurred at the beginning of Experiment 1 when they initially experienced transition from zero proportion of disrupted

communications during Baseline 1 to .33 disrupted communications during Experimental 1. Such an effect, called a schedule contrast effect (Honig, 1966), suggests that accelerated responding during the early intervals of Experimental 1 is at least partially dependent upon the lower proportion of disrupted communications during the preceding Baseline period. Second, the effects of the disrupted communication condition (Experimental 1) appear to extend into the early part of Baseline 2 where responding is depressed at rates below the latter part of that period. Dependency among observations due to the carryover of treatment effects across conditions is discussed by Namboodiri (1972).

2. The rate of communicating by any one person to the other is partially dependent upon the rate of communicating by his or her spouse. For example, the husband would likely not talk to his wife if she did not talk to him and vice versa. The frequency of communications attempted by the husband and wife are said to be cross correlated, because the effects of either person's rate of communicating are linked with concommitant effects on the other person's rate of communicating (Holtzman, 1963). As Figure 1 shows, when the frequency of communications attempted by the husband is high, the frequency of communications attempted by the wife is usually also high and vice versa. However, the cross correlations for the three experimental periods are likely unequal, a circumstance which adds greatly to an already complicated

dependency problem.

3. Repeated measurement of the same subject tends to produce increasing similarity among the subject's scores as the number of trials increases. That is, as the husband and wife acclimate to the experimental situation (i.e., the isolated rooms, the intercom equipment, the "demands" of the situation that they respond in some socially desirable fashion, etc.), each develops a response set. Correlations among observations due to each subject's response set generate a systematic factor which may be labeled "subjective factor" (Shine & Bower, 1971). This dependency among a single subject's scores (i.e. the subjective factor) is further confounded by its possible interaction with particular treatments or particular subjects. It is fairly apparent from looking at Figure 1 that each subject more rapidly develops a response set during each Baseline condition than during the Experimental condition. Thus, there is probably an interaction of the subjective factor with treatment conditions in the data from this prototype experiment.

4. Another type of interobservation dependency, which is related in some respects to the other three dependencies, is that the repeated observations are sequentially related. That is, the rate of communications by either spouse during any given interval is partially dependent upon his rate of communications during preceding intervals. Moreover, his rate of communications during a given interval may also be

affected by any plans he has concerning his behavior during future intervals. When successive observations are sequentially dependent, they are said to be correlated serially (Holtzman, 1963).

These four interobservation dependencies represent the major ways in which errors can be correlated in a repeated measures design with more than one factor. Carryover of treatment effects exemplifies nonindependence between the different levels of one factor and cross correlations between subjects illustrates nonindependence between the levels (assuming that subjects are treated as a factor) of the second factor. Subjective response sets illustrate a type of nonindependence within cells which interacts with the treatment factor, if not also with the subject factor. Serial correlations of successive observations represent nonindependence of observations within cells. The type and the number of interobservation dependencies usually vary with different repeated measures experiments and depend upon the nature of each particular experiment. For example, an experiment involving the administration of multiple dosages of three different drugs to the same subjects could result in treatment carryover effects across the various levels of each drug and numerous nonindependent drug interactions, plus interobservation dependency within cells if multiple observations were obtained on each subject under each condition. This paper now turns to addressing problems encountered when trying to analyze

data of dependent repeated measures with parametric ANOVA techniques.

Problems Encountered with Analysis of Variance

There are three basic assumptions which must be satisfied for valid application of most parametric ANOVA techniques. These assumptions (Hays, 1963) are: (1) the error components of all scores must be normally distributed within cells and across cells; (2) the variance of the error components must be homogeneous across cells; and (3) the error components must be independent within cells and across cells. For heuristic purposes, let us consider what effects the four inter-observation dependencies presented earlier in conjunction with the hypothetical experiment might have with respect to the ANOVA assumptions.

The data of the hypothetical experiment appear to violate all three assumptions underlying parametric ANOVA. In violation of the first assumption, the distribution of error components in Experimental 1 would likely not be normal because of the extreme high scores at the beginning of that period and the extreme low scores toward the end of that period. In violation of the second assumption, the error variance of Experimental 1 would be heterogeneous in comparison to the error variance of Baseline 1 and 2 because of large fluctuations of scores around the mean for Experimental 1. However, violations of these two assumptions, even the quite severe violations here, do not seriously jeopardize the validity of inferences drawn from parametric F test results (Boneau, 1960;

Box, 1953; Lindquist, 1953). It is violations of the third assumption, that the error components are independent, which pose the most serious threat to the validity of the ANOVA results. The effects of the four dependencies are discussed in turn below.

First, carryover of treatment effects from one experimental period to the next would likely diminish differences between the Baseline and Experimental conditions, as well as enhance heterogeneity of error variance. If treatment carryover effects were the only type of dependency operating among the observations, the ANOVA would yield a negatively biased F test for the influence of message disruption on the rate of attempted interspouse communications. Second, positive inter-observation correlations engendered by the dependency of a spouse's rate of communication on the rate of communication by his or her partner would artificially decrease differences between the subjects. This kind of dependency would probably also increase homogeneity of error variance within cells, thus yielding little net effect on the F test for differences between subjects, but contributing to a positively biased F test of differences between experimental conditions. However, grossly unequal cross-correlations for each experimental condition could greatly change this F result and the F test for differences between subjects as well. Third, dependency among observations resulting from each subject's response set characterizes a systematic subjective factor which would tend to increase homogeneity of error variance within each cell and simultaneously would tend to decrease differences between

the experimental conditions. Furthermore, as has been discussed earlier, this troublesome subjective factor appears to interact with treatment conditions. Consequently, it becomes extremely difficult to estimate the net effects of this particular dependency on the various F tests included in the ANOVA, but we might speculate that the F test for differences between subjects would be substantially biased in the positive direction while the F test of treatment effects would probably be less positively biased. Fourth, sequential dependency among the observations would lead to spuriously low error variability, which would result in positively biased F tests for differences between subjects, differences between treatments and their interaction effects.

When all four dependencies among observations are combined, the net effects upon the ANOVA are essentially unknown because the magnitude of each type of dependency relative to the others is extremely difficult to estimate. We know that homogeneity of error variance is artificially increased by three of the dependencies and it is artificially decreased in one case. Furthermore, differences between subjects are artificially reduced in two instances, while effects due to the treatment conditions are reduced in one instance. Overall, nonindependence among the observations suggests that the number of degrees freedom are inflated for each F test. Thus there is a high probability of committing serious errors in inference if the ANOVA is used to test the hypothesis of this

experiment.

In addition to the problems caused by multiple violations of the assumptions underlying the use of parametric ANOVA, there is still another, perhaps equally handicapping, problem with applying either parametric or nonparametric ANOVA to data such as that presented in the prototypic experiment. This problem is that the ANOVA may be statistically insensitive to fluctuations in the data. Statistical insensitivity results from determining error variance by summing deviations of scores within a cell from the cell mean instead of around a trend, such as the apparent linear trends in Baselines 1 and 2 and the apparent curvilinear trends in Experimental 1. When all scores within all cells cluster closely around their respective means, MS error will be small, thus enhancing the possibility of significant F ratios. However, when there is a good deal of error variability around cell means, such as in Experimental 1, MS error will be large and the probability of obtaining significant F ratios will be decreased. Consequently, the ANOVA may be insensitive to important characteristics about the data. Response bursts by both spouses, followed by marked drops in their rates of attempted communication during Experimental 1, are striking examples of important characteristics which are not reflected in the ANOVA application to the data of this hypothetical experiment.

In short, ANOVA is probably not the best statistic for evaluating the data of experiments in which interdependencies

among observations result either by circumstance or design. However, if the experimenter clings to ANOVA as his method of statistical evaluation, there are some alternatives which may enable him to make valid inferences while still following general ANOVA procedures. These alternatives are discussed below.

Solutions Involving ANOVA and Related Techniques

Tests of Assumptions

A preliminary step to applying ANOVA is to determine if the data at hand satisfy the assumptions underlying parametric ANOVA. The assumption of a normal distribution of errors may be tested by following a procedure recommended by Shapiro, Wilk and Chen (1968). The assumption of homogeneity of error variance may be examined by using either Cochran's, Hartley's, or Bartlett's tests (Winer, 1971). In terms of computational complexity, Bartlett's test is most difficult and Winer claims that it does not always yield accurate results, but Gartside's (1972) data do not support this claim. To avoid any doubt, it is best to use the quite sensitive Cochran or Hartley tests.

The assumption of independent errors can not be tested as easily since there are several ways in which nonindependence can be incurred. Evidence bearing on carry-over of treatment effects, subjective response sets, and within-cell serial dependency can be obtained by calculating serial correlations on a single subject's data across all observations and all

treatments. Evidence for nonindependence between subjects can be garnered by calculating cross correlations between paired observations for different subjects. More will be said about these two correlational techniques later.

When the experimental data clearly violate either of the first two assumptions for parametric ANOVA or exhibit any nonrandom correlational pattern, an approximate F test devised by Box (1954) can be employed for tentative hypothesis checking or one of the other solutions involving the ANOVA can be explored. Statisticians might also look into the possibility of applying time series analysis, which will be discussed later.

Corrections for Degrees of Freedom

When any of the values obtained from tests of normality and homogeneity of error variance exceed critical limits, or when the data exhibit only very small interobservation dependencies, it may still be possible to salvage the ANOVA by using a correction for degrees of freedom (df). Greenhouse and Geisser (1959) outline such a procedure. By lowering df, critical values for significance are substantially higher, especially in situations where the n is small. In fact, Winer (1971) contends that the tendency will be to yield too few significant F results.

Nonparametric ANOVA

In some cases where the data show fairly gross departures from normality, it may be acceptable to attempt a nonparametric

ANOVA such as the Kruskal-Wallis (1952) one way ANOVA by ranks or the Friedman (1937, 1940) two-way ANOVA by ranks. Nonparametric ANOVA tests avoid the assumptions of normality and homogeneity of error distributions and tolerate mild violations of the assumption of error independence, but they require ordinal scale measurements such as ranks. Hence, their utility is often limited because the data of most experiments is of the interval type or somewhere between the ordinal and interval levels of measurement (Nunnally, 1970). It is also true that interval scale data can be transformed down to the ordinal scale, but by doing so, there is a loss of information and a greater possibility for Type II errors (McNemar, 1969). Nonparametric statistics, in general, should be used with caution, unless the experiment is designed specifically for ordinal level measurement.

Tests of Means

In some cases where the experimental data fail to satisfy the ANOVA assumptions, it is still possible to redeem the experiment by conducting appropriate a posteriori difference-between-means tests. Most a posteriori tests do not require a set of scores to be broken into component parts such as row, column, interaction and error components, but instead they treat a set of scores within a cell as a single variance. As a result, difference-between-means tests usually detect differences between cells even when the ANOVA error variability is very high compared to variability resulting from

treatments or subjects. Aspin (1949), Cochran and Cox (1957) and Welch (1947) have devised tests for the significance of differences between sets of scores having unequal variances. Consequently, their tests may be useful when the ANOVA assumption of error variance homogeneity is not met. In situations where dependency is incurred by taking single repeated observations of the same subjects under two or more conditions or through paired matching, tests for differences between correlated means can sometimes be conducted (Winer, 1971), but at the cost of reduced df.

When the data depart from normality in fairly gross fashion, it is advisable to explore nonparametric tests for differences between means such as are found in Siegel's (1956) text on nonparametric statistics or most reputable general statistics texts (e.g. Ferguson, 1971; Kirk, 1968; McNemar, 1969). When using nonparametric difference-between-means tests, however, the same precautions apply as were outlined in the preceding section.

There also are some precautions which apply to all difference-between-means tests. Most importantly, while difference-between-means tests are attractive because they maximize the chances of obtaining significant experimental results, this attractiveness should not blind the user to their limitations. Difference-between-means tests do not help the user evaluate comparative treatment effects. Instead, they lump several possible effects together into a single variance estimate for each set of scores. Consequently, difference-between-

means tests are usually poor substitutes for the ANOVA, but they may have some utility if one were interested only in obtaining a crude estimate of treatment or subject effects.

Randomizing Observations

A fifth solution, which may be appropriate in some cases where within-cell dependencies are the only ANOVA violation, is to randomize the observations within each cell (Namboodiri, 1972). In experiments where varying levels of the treatment are sampled and treatment carryover effects are a nuisance, the levels of treatment may sometimes be randomized as well as the observations within each cell. Then, a random effects model, or in some cases a mixed effects model of ANOVA, may be used. Kirk (1968) illustrates the use of a randomized block design when each subject receives all combinations of the treatments. The choice of ANOVA model, of course, rests on the nature of the experimental variables. Most better known texts of the ANOVA (i.e. Hays, 1963; Sheffe, 1959) provide useful guidelines in determining which model to use. An alternative solution of treatment carryover effects is the Latin squares experimental design (Winer, 1971). This solution enables the experimenter to evaluate systematic additive effects such as treatment order effects, but it necessitates the complete randomization or balance of all variables in the experimental design.

The propriety of randomizing the treatments and/or observations within cells is a sticky issue. The decision to randomize or not depends primarily upon how randomization will

affect the tests of the experimental hypotheses (Namboodiri, 1972). If randomization enables the experimenter to test for main effects without artificially biasing the tests for main effects, then randomization is usually permissible. An example of a situation in which it is appropriate to randomize treatments is that situation when order effects may occur, such as an experiment comparing several teaching methods. Here, differences between the teaching methods constitute the main effects and randomization of treatment order will not bias the experimental outcome. Winer's (1971) text is a useful source of advice concerning randomization because it provides extensive information and sample illustrations to help the experimenter determine the propriety of randomization.

Transformations

Another possible solution to interobservation dependencies which is allowable in some situations is transforming the raw scores so as to form a revised distribution which satisfies the assumptions for the ANOVA. A transformation constitutes a change in the scale of measurement. For example, rather than time in seconds (i.e. interval scale data), the revised scale of measurement may be logarithm time in seconds (i.e. ordinal scale data). In experiments where alternative choices of a scale of measurement are theoretically justifiable, the ANOVA of the transformed scores is warranted, provided that the assumptions for ANOVA are satisfied. Whether a transformation of scores is theoretically justifiable or not is determined by the nature of the hypotheses being tested and the scale of

measurement required. If the hypotheses require interval scale data, then transformations which disproportionately change the relative sizes of intervals between scores (e.g. a logarithmic transformation) are not permissible. However, if the hypotheses require only ordinal scale data, then monotonic transformations which do not disturb the order of the scores are theoretically justifiable (e.g., square root, logarithmic and arc-sine transformations). An example of a situation in which transformation of scores is theoretically justifiable is an experiment comparing different teachers' judgments about several groups of students on seven-point Likert scales. Here, the data is of the ordinal level and monotonic transformation of the raw scores will probably not affect the hypotheses which the experimenter is testing. Kirk (1968) and Winer (1971) outline methods of transforming raw scores and they discuss the occasions when each is an allowable operation.

Multivariate ANOVA

In experimental designs entailing multiple observations of the same subjects on two or more dimensions, it is appropriate to use multivariate ANOVA techniques. It may be reasonable in some instances to apply multivariate ANOVA to cross correlated observations between subjects simply by assuming that the different subjects are different dimensions of the same unit. Morrison (1967) and Overall and Klett (1972) have provided two useful texts on multivariate ANOVA.

But, within-subject dependencies of repeated measures experiments pose more complicated problems for multivariate

ANOVA techniques. In order to deal with problems, Finn (1969) and Poor (1973) have proposed an alternative multivariate ANOVA for repeated measures designs which eliminates the within-subject factors from the experimental design by defining a new set of transformed variables representative of the within-subject effects. Since the mathematical calculations associated with this alternative ANOVA technique are available elsewhere (i.e. Finn, 1969; Poor, 1973), they will not be presented here. But, a brief discussion of the general approach seems warranted.

The multivariate method devised by Finn and Poor treats all replications for each subject within a cell as added factors. For example, if there are five observations on each subject within a cell, there are five added factors. The raw scores within a cell are then transformed to form a revised set of five additional weighted factors. Each new weighted factor is evaluated by testing the null hypothesis that it has no effect. Using the Finn and Poor procedure greatly increases the number of F tests which must be calculated, but the procedure enables the user to evaluate data with intra-subject dependencies.

Other Remarks Concerning ANOVA

The statistics discussed thus far in this paper apply only to experimental designs with equal cell frequencies. When the number of observations per cell are not equal, the effect is to produce inaccurate probability statements concerning the ANOVA results. This problem may be handled by a

least squares solution for unequal cell frequencies (Overall & Spiegel, 1969, 1973; Rawlings, 1972, 1973) or by calculating expected proportionate cell frequencies (Tsao, 1946; Bancroft, 1968).

It has been assumed throughout this discussion thus far that dependency among repeated observations should not necessarily be viewed as a nuisance to be minimized or controlled for in experiments. If the experimenter takes the position that interobservation dependencies are a detraction to the results, then he should take precautions, such as Namboodiri (1972) outlines, in designing his study. In many experiments though, dependencies among the observations constitute some of the more interesting and theoretically important features of the data.

Alternative Statistics: Trend Analysis

Analyses for trends and differences between trends are useful in certain single and multiple factor experiments where repeated observations are made of the same subject(s). Trend analysis can be viewed as an extension of the ANOVA in that it enables the user to evaluate more precisely the nature of relationships between variables. Instead of determining only if manipulating the independent variable produces significant effects in the dependent variable (as the ANOVA does), trend analysis also allows the user to test for the significance of various kinds of relationships (i.e. linear, quadratic, exponential relationships, etc.) between the independent and dependent variables.[3]

As with ANOVA, there are a number of shortcomings of trend analysis when it is applied to data containing one or more interobservation dependencies. The principal shortcomings of trend analysis ensue from its underlying assumptions. Since trend analysis is an extension of the ANOVA, its primary underlying assumptions are the same, namely, a normal distribution of error, homogeneity of error variance and error independence (Alexander, 1946; Grant, 1956). If we were to consider applying trend analysis to the data of the hypothetical experiment presented near the beginning of this paper, we would find that all three assumptions are violated multiply by the data. These violations are the same as were discussed in relation to the ANOVA.

An additional assumption of trend analysis, which is not essential to the ANOVA, is that the levels of an independent variable must be equally spaced to permit use of the usual (i.e. Ferguson, 1971; Fisher & Yates, 1953) tables of orthogonal polynomials. For example, if one were investigating the effects of the popular sedative chloral hydrate on blood pressure, several equally spaced dosages (e.g. 0 mg., 200 mg., 400 mg., 600 mg., 800 mg.) of the independent variable might be used. When the levels of the independent variables are unequal but monotonically ordered, Robson's (1959) modification for constructing orthogonal polynomials may be employed. Although the treatment conditions of the independent variable in the exemplary experiment are apparently equally spaced (i.e., 0 messages disrupted, .33 messages disrupted and 0

messages disrupted), carryover of treatment effects from one condition to the next tends to truncate the steps between conditions and perhaps even to introduce inequality between the steps. Since the magnitude of each treatment carryover effect is impossible to estimate accurately in this example, Robson's modification for constructing orthogonal polynomials has no utility here. Moreover, the treatments are not monotonically ordered, a feature which makes the nature of the relationship between the independent variable and the dependent variable substantially more difficult to estimate.

Besides the shortcomings caused by multiple violations of the assumptions underlying trend analysis, there is still another perplexing problem with applying trend analysis to data such as that of the hypothetical experiment. Just as with the ANOVA, there is a problem of statistical insensitivity to data fluctuations within each treatment condition. This insensitivity is incurred in two ways. First, a large error variance results from marked fluctuation of scores within certain cells around their respective means (as in Experimental 1). When error variance is large compared to trend effects, such as in this hypothetical case, the chances of finding a significant relationship between the independent and dependent variables is decreased. Second, when the mean frequencies of attempted communication per conditions are graphed, the means appear to fit a straight line with zero slope. As in any regression analysis, the finding of zero slope suggests no relationship between two variables (Ghiselli, 1964). Thus, while trend

analysis would show no relationship between the treatment conditions and the frequency of attempted communication, it is nonetheless obvious that changes in the treatment are responsible for several notable changes in the frequency of attempted communication within cells during Experimental 1 and Baseline 2.

In sum, application of trend analysis to experimental data containing one or more interobservation dependencies may be precluded by violations of underlying assumptions and by statistical insensitivity to important characteristics of the data. There are, however, several alternative procedures involving trend analysis which experimenters might explore. For the most part, these alternatives parallel those presented in relation to the ANOVA, but there are enough differences to warrant their resume.

Tests of Assumptions

Whenever there is any reason to suspect interobservation dependency or when the experimenter has no knowledge about the effects of the treatments upon the variance, it is wise to test the assumptions before applying trend analysis. The same techniques are used as with ANOVA. Moreover, trend analysis, like ANOVA, can withstand fairly gross departures from normality and homogeneity in the error distributions, but only very small nonrandom interobservation correlations. Even if the assumptions are clearly violated, it is sometimes helpful to calculate expected mean squares in order to see if there are any potentially interesting effects which might be

scrutinized further with other correct alternative statistics.

Corrections for Degrees of Freedom

As with ANOVA, the Greenhouse and Geisser (1959) method of estimating conservative df can be used to test for main effects, interaction effects, regression effects, and deviations from regression. Reducing df also diminishes statistical sensitivity however.

Nonparametric Trend Analysis

Ferguson (1965) has presented methods of nonparametric trend analyses which have less stringent requirements for their application than their parametric counterparts. Nonparametric trend analyses avoid the usual assumptions of error normality and homogeneity and tolerate small degrees of dependence. They also do not require equal spaces between levels of the independent variable. However, because nonparametric statistics use ordinal scale measures (i.e. ranks), their use is limited to experiments designed specifically for ordinal scale measurement. Inasmuch as ordinal scale measures are used, the possible trend relationships between an independent and dependent variable which may be tested for significance are necessarily monotonic, bitonic (i.e. shaped), and so on, rather than linear, quadratic, and so forth.

Hotelling T^2

Hotelling's T^2 is a multivariate generalization of the univariate Student's t which appears to have utility in some repeated measures designs for detecting a trend associated

with varying levels of treatment (Overall & Klett, 1972). A noteworthy advantage of the $\underline{T}^2$ statistic is that it does not require equal correlations among the observations. $\underline{T}^2$ tests the hypothesis that $M_1 = M_2 = M_3 \ldots = M_j$ where M_1 is the mean of all scores in cell 1, M_2 is the mean of all scores in cell 2 etc. With regard to its utility in repeated measures designs, the use of $\underline{T}^2$ is limited to single factor experiments where the same subjects are measured on several levels of the same treatment. $\underline{T}^2$ enables the experimenter to determine if there is a trend associated with the treatment levels which differs significantly from zero slope, but it does not enable the user to specify the nature of the trend (i.e. whether it is linear, quadratic etc.) or the magnitude of the trend. Consequently, $\underline{T}^2$ should be used only as a pretest for general trend when observations are correlated unequally. The statistical mechanics of the $\underline{T}^2$ test are available in texts by Overall and Klett (1972) and Winer (1971).

Replications as a Factor

A solution involving trend analysis which may be appropriate when the only interobservation dependency is correlated scores within cells, is to treat replications as an added factor. Such a situation might arise when the same subjects are measured repeatedly on the same treatment condition. Thus, if treatments were the only factor (as in the usual one-way trend analysis), replications would be a second factor and the scores would be arranged into a two-way ANOVA matrix

with one score per cell. If the experiment called for a two-way trend analysis with repeated measures within each cell, the repeated measures would be treated as a third factor and the scores would be arranged into a three-way ANOVA matrix with one score per cell. McNemar (1969) has outlined special cases of linear and quadratic trend analyses and analyses for differences between trends for one and two factor experiments in which the subjects are measured repeatedly. McNemar also has presented several special cases of analyses for trend differences in which the observations are correlated across columns or across both columns and rows. In order to apply these special trend analyses, it is essential that the observations are equally spaced in time.

Transformations

As with the ANOVA, sometimes the assumptions underlying trend analysis can be satisfied by transforming the raw scores into a revised distribution. Bobbit, Gourevitch, Miller and Jensen (1969), for example, used the arc-sine method of transforming raw data consisting of proportions into a revised distribution which satisfied the assumptions for trend analysis. Since a score transformation changes the scale of measurement, it is necessary that the transformation be theoretically justifiable. The cautions concerning the theoretical justification of transforming scores, which were outlined in relation to the ANOVA earlier in this paper, apply to trend analysis as well.

Other Remarks about Trend Analysis

The trend analyses discussed in this paper have all presumed equal score frequencies per cell. Equal cell frequencies make the task of constructing orthogonal polynomials easier. When cell frequencies are unequal, least squares adjustments must be made. These adjustments can be quite complicated, but they are aptly illustrated by Ferguson (1971).

A few summary conclusions about trend analysis are warranted before moving on to other statistical techniques for handling data involving repeated measurements of the same subjects. First, in situations where the assumptions of both ANOVA and trend analysis are satisfied and the data within cells cluster closely around a cell mean, then trend analysis is a better choice of statistical analysis because of its greater precision in clarifying the relationship between independent and dependent variables. Second, trend analysis, like the ANOVA, is severely limited in its applications to experiments where repeated measures of the same subjects are taken, because numerous sources of interobservation correlation may invalidate the technique. Third, experimenters should carefully scrutinize and test data to insure that it is best suited to trend analysis or to determine if the data is handled better by other statistical techniques.

Alternative Statistics: Time Series Analysis

Time series analysis, it is suggested here, may be the optimal statistic for evaluating the data of many experiments

involving repeated measures of the same subjects. Time series analysis is particularly well suited to repeated measures designs because it makes positive use of interobservation dependencies by generating functions which fit the dependencies (Box & Jenkins, 1970). Gottman, McFall and Barnett (1969) observe that "the dependent time series can be understood as consisting of a signal (the underlying dependency of the observations over time) which has been combined with white noise (error). The generating function, then operates to separate the signal from the noise" (p. 302).

Four models of generating functions have been proposed for use by social science researchers and statisticians who are looking to time series analysis as a way of analyzing dependent repeated measures data. They are the autoregressive model (AR), the moving average model(MA), the autoregressive moving average model (ARMA) and the autoregressive integrated moving average model (ARIMA). The autoregressive integrated moving average model holds that the value of an observation in a time series can be estimated best by taking a statistically-derived constant times the previous observation, plus an unpredictable error term (Gottman, McFall & Barnett, 1969). The moving average model holds that the value of an observation in a time series is given by an "exponentially weighted moving average" of previous observations, plus an unpredictable error term. This means that an observation remembers the previous observation most and the other preceding observations less and less according to how far they precede the given observations

(Gottman, McFall & Barnett, 1969). The ARMA and ARIMA models contain elements of both the AR and MA models (Box & Jenkins, 1970). The basic task of time series analysis is to find which model best fits a given series of observations.

The four time series models proposed here may only be used when the data consist of discrete observations. When the data consist of continuous or nearly continuous observations, spectral analysis is used. Spectral analysis and its applications are discussed in detail by Jenkins and Watts (1968). Within psychology and other related social sciences however, most data are comprised of single or multiple series of discrete observations. Series of discrete measures taken sequentially over time may arise in two ways: (1) Sampling a continuous time series; for example, obtaining daily measures of an astronaut's blood sugar level while he is on a flight to and from the earth's moon; (2) Accumulating a variable over a period of time; for example, obtaining cumulative bar presses of an animal on a fixed interval reinforcement schedule (Box & Jenkins, 1970). Generally, future values of these discrete time series cannot be predicted exactly in the social sciences, hence the term "stochastic" (i.e. probability) processes is used to describe these series.

Campbell (1963), Campbell and Stanley (1966), Gottman (1973) and Glass, Willson and Gottman (1973) have described a number of repeated measures designs to which time series analysis is amenable. Glass et al. discuss the application of

time series analysis to these designs and potential sources of invalidity in time series experiments. For example, the experiment presented at the beginning of this paper fits the "multiple-group--multiple-intervention" time series experimental design described by Glass et al. This design entails two or more different interventions which occur simultaneously in two or more time series. In the hypothetical experiment the husband and wife are the two "groups" in this case. The multiple interventions here consist of the change from Baseline 1 to Experimental 1 and the change from Experimental 1 to Baseline 2 conditions. These changes occur simultaneously for both subjects. Essentially then, each time series is a replication of the other. According to Glass et al., the "multiple-group--multiple-intervention" design enjoys a large measure of validity because the "groups" are matched over time and differences between "groups" can be ascertained if any occur. The chance that the differences in intervention effects are the result of historical accidents coincident with the interventions is slight. Nonetheless, this possibility could be reduced even further if the subjects were studied for a greater duration of time in each condition and if further reversals of treatment conditions were undertaken.

It is not the intent of this paper to review all of the work of Glass et al. (1973) or to discuss all the possible applications of time series analysis to dependent repeated measures data. Rather, it is deemed that the interests of the reader may be best served by illustrating the necessary steps

involved in a typical time series data analysis.

Stages in Model Building

The basic problem of finding a function which fits a given series of observations can be divided into three stages: (1) Identifying the model which best fits each series by specifying its type (i.e. AR, MA, ARMA, or ARIMA), order (i.e. linear, quadratic, etc.) and whether it is stationary or nonstationary (i.e., whether the series oscillates around a mean or has no natural mean); (2) Estimating the parameters of the model from the observed data; and (3) Diagnostic checking of the model's goodness-of-fit by testing the residuals of the fitted model from the observed data for randomness (Gottman, McFall & Barnett, 1969).

1. Identifying the model which best fits a given time series begins with calculating autocorrelation and partial autocorrelation functions on the observed data. Autocorrelation and partial autocorrelation are the tools for estimating the amount of predictable interobservation dependency. They also help determine the type and order of the best-fit time series model and they indicate whether the raw data are stationary or nonstationary.

The autocorrelation function is defined as the correlation of observations over time, obtained by pairing observations k intervals apart (k = 1,2,...). This gives the serial correlations as a function of lag. The formula for obtaining each serial correlation r_k, as given by Anderson (1971), is

$r_k = \frac{\Sigma x_t x_{t+k}}{x^2_t}$ where the x are observations in a series $x_1, x_2 \ldots xT$. This formula is similar to the Pearson product-moment correlation coefficient. When the serial correlations are plotted as a function of lag, the resulting figure is called a correlogram. Comparison of the obtained correlogram with known correlograms for the basic AR, MA, ARMA, and ARIMA models helps determine which model best fits the raw data. Glass et al. (1973, p. 88) present the expected correlograms for first and second order AR and MA models and the mixed first order autoregressive and moving average model. Figure 2 illustrates exemplary correlograms of a pure first order autoregressive model with $r_{lag = 1} = -.7$ and a pure first order moving average model with $r_{lag = 1} = .8$. As can be seen, the pure first order autoregressive model contains serial correla-

---

Insert Figure 2 about here

---

tions which gradually approach zero as successive powers of the first serial correlation. The correlogram for a pure first order moving average model shows an abrupt drop to zero after a lag of 1. For mixed models (i.e. ARMA and ARIMA), the autocorrelation function dies out in an exponential manner similar to the autoregressive model.

The significance of a serial correlation may be tested by using R. L. Anderson's (1942) exact method of determining statistical significance. When none of the serial correla-

tions is significant, the series is said to be independent of time and the observations may be treated as one would any other set of independent observations. However, the serial correlations become less stable as the lag increases and the number of pairs of observations shrinks. Thus, the last few serial correlations for a given series of observations may be considered meaningless.

The partial autocorrelation function is used to help confirm which model best fits the raw data, but its principal use is to specify the order of the best-fit model. The partial autocorrelation function consists of partial serial correlations which are usually arranged in a correlogram so as to indicate the number of preceding observations necessary to account for the time determined dependency of the raw data series. Partial serial correlations can be obtained by solving Yule-Walker equations (Glass et al., 1973, p. 85-87) or by following the simpler procedures outlined by Jenkins and Watts (1968) and Quenouille (1957). As Quenouille has shown the first partial serial correlation $\hat{\pi}_1$ is always approximately the same as $r_{lag = 1}$. When $\hat{\pi}_1$ is significant and all other $\hat{\pi}$ are insignificant, the best-fit time series model is linear. The second partial serial correlation, $\hat{\pi}_2$ is approximated by the formula $\hat{\pi}_2 = \frac{r_{lag = 1} - r_{lag = 2}}{1 - r^2_{lag = 2}}$ and it indicates excess correlation which is not expressed in a first order model. In other words, if $\hat{\pi}_2$ is significant

and subsequent $\hat{\pi}$ are insignificant, the best-fit model is probably a second order (i.e. quadratic) process. In a similar manner, higher order partial serial correlations can be computed. Usually, according to Glass et al., most data are adequately described by a first or second order model. Although it is often possible to "fit" higher order models to many data series, there is a real danger of overfitting to the point where artificial correlation may be induced into the data.

The situation frequently arises in the social sciences in which the autocorrelation and partial autocorrelation functions calculated on the raw data do not fit the expected patterns of any of the typical models presented by Glass et al. (1973, p. 88). If the serial correlations or partial serial correlations remain large or unstable over many lags instead of fitting one of the expected patterns, then the raw data series can be considered nonstationary and steps should be taken to produce a stationary series with which to work. A stationary series can be achieved by calculating successive differences on the raw data until a series is obtained which fits one of the expected models in terms of its autocorrelation and partial autocorrelation functions. To illustrate, first differences are calculated by subtracting observations x_2 from x_1, then x_3 from x_2 and so forth. Second differences are obtained by squaring the first differences.

In summary, the first stage of time series model building

involves calculating autocorrelation and partial autocorrelation functions on the raw data series or on successive differences of the raw data until stationary correlograms are obtained which fit expected patterns typical of either the AR, MA or mixed models. A pure first order AR process is characterized by serial correlations which die out exponentially as powers of $r_{lag = 1}$ and partial serial correlations which approach zero after lag = 1. A second order AR process has a similar autocorrelation correlogram but the partial autocorrelation function cuts off abruptly after lag = 2. A pure first order MA process has an autocorrelation function which approaches zero after lag = 1, but the partial autocorrelation function dies out more gradually after a significant $\hat{\pi}_1$. The second order MA process has a similar shaped autocorrelation correlogram but $\hat{\pi}_2$ is significant. For mixed models both the autocorrelation and partial autocorrelation functions die out exponentially.

2. After the type, order and degree of differencing of the time series model have been identified, the parameters of the model must be estimated from the observed data. Stated mathematically, the task is to estimate the value of F (B) in the time series $z_t - L = F(B)a_t$ which minimizes a_t where: z_t is any point in the observed time series, L is the mean of that series, a_t is error, and F(B) is the function of a backward shift operator (i.e. a "signal" consisting of predictable interobservation dependency). In other words, this step involves generating a function which

minimizes differences between the generated model and the observed data series (Glass et al., 1973). The F in the equation consists of weights which are symbolized by ϕ for autoregressive processes and θ for moving average processes and B indicates the backward shift operator (e.g. t-1, t-2, etc.). By substitution, the equation for a first order AR process is $z_t - L = \phi_1 z_{t-1} + a_t$; the equation for a second order AR process is $z_t - L = \phi_1 z_{t-1} + \phi_2 z_{t-2} + a_t$; The equation for a first order MA process is $z_t - L = a_t - \theta_1 a_{t-1}$; the equation for a second order MA process is $z_t - L = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2}$ and the equation for a first order mixed process if $z_t - L - \phi_1 z_{t-1} = a_t - \theta_1 a_{t-1}$ (Box & Jenkins, 1970).

With a little algebra (see for example, Box & Jenkins, 1970, Chapter 7), it can be shown that the total errors of "fit" $\Sigma \hat{a}_1^2$ vary as a function of varying ϕ and θ, that is $\underline{SS}(\phi, \theta) = \sum_{i=1}^{N} \hat{a}_1^2$. In practice, the values of ϕ and θ which minimize Σa_1^2 can be found by setting ϕ and θ equal to values between -1.0 and 1.0 and calculating $\underline{SS} = \Sigma a_i^2$ for all observations in the series. The minimum $\underline{SS}$ determines the maximum likelihood ϕ_1 and θ_1. A confidence interval around the generated model can be calculated by following procedures outlined by Glass et al. (1973, p. 90). Calculating $\Sigma \hat{a}_i^2$ for varying values of ϕ and θ can be laborious if it is done by hand, so it may be wise to take advantage of available computer programs by Box and Jenkins (1970) and Gottman (1973) which

perform these calculations.

3. Diagnostic checking of the goodness-of-fit of the generated model can be done by computing serial correlations on the residuals $\hat{a}_t = z_t - \hat{z}_t$. If the correct model has been fitted to the raw series, the resulting autocorrelation function should exhibit a pattern of random serial correlations. However, if the residuals exhibit an autocorrelation pattern similar to a known model, then the model should be fit to the residuals. This procedure is equivalent to fitting a mixed model to the original data (Glass et al., 1973).

Significance Testing

Glass et al. (1973) discuss in depth how to test for the significance of unplanned interventions (e.g. accidents) and planned interventions (e.g. experimental treatments) in series' of repeated observations on the same subjects. Since most experimental research entails making planned interventions, this section is devoted to testing the significance of treatment effects in time series experimental designs.

The procedure for testing the significance of a treatment effect in time series designs is quite simple because it is assumed that a model generated for a data series under one condition (e.g. baseline) holds for another condition (e.g. experimental condition). The sum of squares of residuals between the observed data of the experimental condition and the previously generated model for the baseline condition are calculated according to the formula $SS_1 = \Sigma(x_t - x_t)^2$

where $\hat{x}_t$ is the forecasted value of x_t. Then the best-fit model for the data of the experimental condition is generated and the sum of squares of residuals (SS_2) between the generated model and the observed series are calculated. Finally, a test of statistical significance is computed by comparing the SS residuals of the forecasted model with the SS residuals of the generated model according to the formula

$$F = \frac{(SS_1 - SS_2)/df_1}{(SS_1 + SS_2)/df_2}$$ where: df_1 is the number of parameters involved in the model estimation (i.e. two in this case) and $df_2 = N_2 - 2$ where N_2 is the number of observations during the experimental condition (Gottman, McFall & Barnett, 1969).

It should be mentioned that time series analysis is quite sensitive to detecting shifts in the data of a single subject within a series because that function is found which minimizes the error around a statistically-derived trend. It is certainly more sensitive to shifts within a series than the ANOVA and trend analysis, which estimate error around a cell (series) mean. When unplanned shifts in trend within a single series (cell) occur, the Glass et al. (1973) method can be followed for determining the point where the shift began.

Examining Cross Correlated Series

The time series analyses discussed thus far account for three of the types of interobservation dependency presented at the outset of this article, namely, treatment carryover effects, response set effects and sequential dependency. These three interobservation dependencies operate within a single

subject. The question arises, however, as to what extent a single subject's response data is influenced by concommitant variation in the responses of another subject or dependent variable of the same experiment. In other words, how can cross correlated time series be examined?

In order to investigate the amount of relationship among two or more matched dependent variables, product-moment correlation coefficients (i.e. cross correlations) may be computed and tested for significance if the serial correlations of each series are insignificant. If any of the serial correlations of any of the series being examined are significant however, the ordinary tests of significance of the cross correlations among the series must be adjusted. The adjustment of the significance test proceeds by estimating the number of "effective independent observations" ($\underline{N}'$) according to Bartlett's (1935) formula $N' = N/(1 + 2r_1r'_1 + 2r_2r'_2 + \ldots)$, where $\underline{N}$ is the number of pairs on which the usual r_{xy} is based, r_1, $r_2 \ldots$ are the serial correlations for the series $\underline{x}$, and r'_1, $r'_2 \ldots$ are the corresponding serial correlations for series $\underline{y}$. The degrees of freedom for testing the significance of the cross correlation are then equal to $N' - 2$. As Gottman (1973) and Holtzman (1963) indicate, cross correlations among time series can also be computed by systematically lagging $\underline{x}$ behind $\underline{y}$ and vice versa by $k = 1, 2, \ldots$

If the cross correlation between two series is statistically significant, the relationship between the two series may be examined further by attempting to fit the model generated

for series $\underline{x}$ to series $\underline{y}$ and vice versa. This technique is similar to Gottman's (1973) use of transfer functions. Another procedure which this author suggests is to attempt to fit a model to the residuals (d_1, d_2 ...d_T) between $\underline{x}$ and $\underline{y}$ where $d_1 = x_1 - y_1$, $d_2 = x_2 - y_2$ etc. The generated model which best fits the residuals may help clarify the relationship between the two raw data series. For example, if the generated model which best fits the residuals $\underline{d}$ is a first order MA process, this means that variation between the two series $\underline{x}$ and $\underline{y}$ is predictable as a function of a moving average of exponentially weighted previous observations.

There are several precautions which apply to the interpretation of relationships among series of dependent observations. First, finding significant cross correlation (lagged or unlagged) between two series and successful fitting of the same generated model to both series or of a new model to the residuals between the matched observations may warrant any of several conclusions. It could mean that variation in series $\underline{x}$ causes variation in series $\underline{y}$ or vice versa, or that series $\underline{x}$ and $\underline{y}$ are similar processes, or that series $\underline{x}$ and $\underline{y}$ have a similar cause. Only when these analyses yield negative results can a more conclusive statement about causal inference be made. If two series are uncorrelated, it is unlikely that they are causally connected (Gottman, 1973). Second, further experimental research is needed in which the variables $\underline{x}$ and $\underline{y}$ are systematically manipulated as independent variables and the consequent effects of one upon the other are studied.

Moreover, any other likely hypotheses for the concommitant variation in x and y (e.g. similar history, accidents) must be eliminated. As Gottman (1973) suggests, "causal connection is never demonstrated; rather, we successively eliminate rival hypotheses which militate against confidence in causal connection" (p. 99). Third, the use of cross correlation, transfer functions and generating models which fit the residuals between two series are at the exploratory stage in social science research. Consequently, the application of these statistics must be subjected to much further investigation and review to confer greater credibility upon them as viable means of analyzing cross correlated data.

Several comments about the application of time series analysis in general to repeated measures data are also warranted at this point. There is a good deal of precedence for the use of time series analysis within economics and engineering but its application within psychology and related social sciences has been more recent. Consequently, the use of time series analysis in the social sciences should be considered carefully and its implications should be examined as thoroughly as possible. The diligent statistician who attempts to fulfill these conditions should be prepared for time-consuming work, but the aforementioned computer programs may facilitate his efforts. Finally, one problem which requires particular further clarification in social science research is that of determining the most appropriate interval duration for a given observation. Generally, the determination of interval duration in time

series experiments is arbitrarily made on the basis of convenience (e.g. a day), but it can be argued that there should also be statistical justification for selecting a certain interval duration. That is, the interval should be short enough to sensitively reflect real changes in the dependent variable (e.g. behavior), but long enough to allow for reasonable stability in response rates if the data really does have a stationary component.

Other Alternatives

This presentation of ANOVA, trend analysis and time series analysis has by no means exhausted all alternative means of examining dependent repeated measures data. To be sure, graphical illustration of data may be sufficient in some situations where an interpretation is obvious simply by eyeballing the raw data and no further clarification of the results is desired. However, there seems to be a trend among journal editors these days toward accepting studies for publication which include statistical tests of data, even when interpretations of the data appear clear from graphical illustrations alone.

A simple technique which may be useful for evaluating the association between two variables sampled over time (i.e. multiple time series) is the coefficient of directional correlation (Strahan, 1971). Suppose that two time series x and y consist of matched observations. The coefficient of directional correlation (d) estimates the tendency for the two series to change values in the same direction from one

sampling occasion to the next. It is computed according to the formula $d = \frac{p - m}{p + m}$ where: p is the number of instances in which the direction of x is the same as that of y for matched intervals (i.e. both variables increase or decrease), and m is the number of instances in which the direction of x is opposite that of y (i.e. one variable increases and the other decreases).

There is some debate (e.g. Holtzman, 1963) whether or not the value of d is inflated when serial correlation is present in one or both series being compared. However, Cattell (1963) argues that since the d technique is a nonparametric statistic, the presence of serial correlation makes little difference. Wiener (1949) has devised a more sophisticated procedure for calculating a correlation coefficient between time series which avoids concerns raised about the d technique by providing information about both the signal (i.e. predictable interobservation dependency) and noise components of each series.

Summary

This paper has shown several ways in which interobservation dependencies are incurred when repeated measures are made on the same subjects. The presence of significant interobservation dependencies may invalidate the use of ANOVA and its extension, trend analysis, as statistical techniques for analyzing the data of dependent repeated measures. Tests were presented for determining whether the assumptions underlying the use of the parametric ANOVA and trend analysis are

violated. Several alternative statistical procedures which involve the ANOVA and trend analysis, and which obviate problems of interobservation dependencies, were discussed.

However, it is suggested that time series analysis may be the optimal statistic for evaluating the data of many experiments involving repeated measures of the same subjects because time series analysis makes positive use of interobservation dependencies by generating functions which fit the dependencies. Moreover, because time series analysis entails calculating error around the best-fit trend within a single series (i.e. cell) instead of around a cell mean (as in the ANOVA and trend analysis), it tends to be more sensitive to shifts in the data of a single subject within a cell than either the ANOVA or trend analysis. The procedures for conducting a typical time series analysis were presented and several precautions associated with applying this statistic were also outlined. Finally, the advantages and limitations of graphic display only of data and simple directional correlation were briefly discussed as alternatives to the rather time-consuming procedures of time series analysis.

References

Alexander, H. W. A general test for trend. Psychological Bulletin, 1946, 43, 533-557.

Anderson, R. L. The distribution of the serial correlation coefficient. Annals of Mathematical Statistics, 1942, 13, 1-13.

Anderson, T. W. The statistical analysis of time series. New York: Wiley, 1971.

Aspin, A. Tables for use in comparisons whose accuracy involves two variances separately estimated. Biometrika, 1949, 36, 290-291.

Bancroft, T. A. Topics in intermediate statistical methods. Ames, Iowa: Iowa State University Press, 1968.

Bartlett, M. S. Some aspects of the time-correlation problem in regard to tests of significance. Journal of the Royal Statistical Society, 1935, 98, 536-543.

Bobbitt, R. A., Gourevitch, V.P., Miller, L. E. & Jensen, G. D. Dynamics of social interactive behavior: A computerized procedure for analyzing trends, patterns, and sequences. Psychological Bulletin, 1969, 71, 110-121.

Boneau, C. A. The effectx of violations of assumptions underlying the t test. Psychological Bulletin, 1960, 57, 49-64.

Box, G. E. P. Non-normality and tests on variance. Biometrika, 1953, 40, 318-335.

Box, G. E. P. Some theorems on quadratic forms applied in the study of analysis of variance problems. Annals of Mathematical Statistics, 1954, 25, 290-302.

Box, G. E. P. & Jenkins, G. M. Time series analysis: Forecasting and control. San Francisco: Holden-Day, 1970.

Campbell, D. T. From description to experimentation: Interpreting trends as quasi-experiments. In C. W. Harris (Ed.), Problems in measuring change. Madison, Wisconsin: University of Wisconsin Press, 1963, Pp. 212-242.

Campbell, D. T. & Stanley, J. C. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1966.

Cattell, R. B. The structuring of change by P-technique and incremental R-technique. In C. W. Harris (Ed.), Problems in measuring change. Madison, Wisconsin: University of Wisconsin Press, 1963. Pp. 167-198.

Cochran, W. G. & Cox, G. M. Experimental designs. (2nd ed.) New York: Wiley, 1957.

Dodd, D. H. & Schultz, R. F., Jr. Computational procedures for estimating magnitude of effect for some analyses of variance designs. Psychological Bulletin, 1973, 79, 391-395.

Ferguson, G. A. Nonparametric trend analysis. Montreal: McGill University Press, 1965.

Ferguson, G. A. Statistical analysis in psychology and education. (3rd ed.) New York: McGraw-Hill, 1971.

Finn, J. D. Multivariate analysis of repeated measures data. Multivariate Behavioral Research, 1969, 4, 391-413.

Fisher, R. A. & Yates, F. Statistical tables for biological, agricultural, and medical research. (4th ed.) Edinburgh and London: Oliver & Boyd, 1953.

Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. Journal of the American Statistical Association, 1937, 32, 675-701.

Friedman, M. A comparison of alternative tests of significance for the problem of m rankings. Annals of Mathematical Statistics, 1940, 11, 86-92.

Gartside, P. S. A study of methods for comparing several variances. Journal of the American Statistical Association, 1972, 67, 342-346.

Ghiselli, E. E. Theory of psychological measurement. New York: McGraw-Hill, 1964

Glass, G. V., Willson, V. L. & Gottman, J. M. Design and analysis of time-series experiments. Boulder, Colorado: Laboratory of Educational Research, 1973.

Gottman, J. M. N-of-one and n-of-two research in psychotherapy. Psychological Bulletin, 1973, 80, 93-105.

Gottman, J. M., McFall, R. M. & Barnett, J. T. Design and analysis of research using time series. Psychological Bulletin, 1969, 72, 299-306.

Grant, D. A. Analysis-of-variance tests in the analysis and comparison of curves. Psychological Bulletin, 1956, 53, 141-154.

Greenhouse, S. W. & Geisser, S. On methods of analysis of profile data. Psychometrika, 1959, 24, 95-112.

Hays, W. L. Statistics for psychologists. New York: Holt, Rinehart & Winston, 1963.

Holtzman, W. H. Statistical models for the study of change in the single case. In C. W. Harris (Ed.), Problems in measuring change. Madison, Wisconsin: University of Wisconsin Press, 1963. Pp. 199-211.

Honig, W. K. (Ed.) Operant behavior: Areas of research and application. New York: Appleton-Century-Crofts, 1966.

Jenkins, G. M. & Watts, D. G. Spectral analysis and its applications. San Francisco: Holden-Day, 1968.

Kirk, R. E. Experimental design: Procedures for the behavioral sciences. Belmont, California: Brooks-Cole, 1968.

Kruskal, W. H. & Wallis, W. A. Use of ranks in one-criterion variance analysis. Journal of the American Statistical Association, 1952, 47, 583-621.

Kurtz, K. H. Foundations of psychological research. Boston: Allyn & Bacon, 1965.

Lindquist, E. F. Design and analysis of experiments in psychology and education. Boston: Houghton Mifflin, 1953.

McNemar, Q. Psychological statistics. (4th ed.) New York: Wiley, 1969.

Namboodiri, N. K. Experimental designs in which each subject is used repeatedly. Psychological Bulletin, 1972, 77, 54-64.

Nunnally, J. C., Jr. Introduction to psychological measurement. New York: McGraw-Hill, 1970.

Overall, J. E. & Klett, C. J. Applied multivariate analysis. New York: McGraw-Hill, 1972.

Poor, D. D. S. Analysis of variance for repeated measures designs: Two approaches. Psychological Bulletin, 1973, 80, 204-209.

Quenouille, M. H. The analysis of multiple time series. New York: Hafner, 1957.

Rawlings, R. R., Jr. Note on nonorthogonal analysis of variance. Psychological Bulletin, 1972, 77, 373-374.

Rawlings, R. R., Jr. Comments on the Overall and Spiegel paper. Psychological Bulletin, 1973, 79, 168-169.

Robson, D. S. A simple method for construction of orthogonal polynomials when the independent variable is unequally spaced. Biometrics, 1959, 15, 187-191.

Shapiro, S. S., Wilk, M. B., & Chen, H. J. A comparative study of various tests for normality. Journal of the American Statistical Association, 1968, 63, 1343-1373.

Sheffe, H. The analysis of variance. New York: Wiley, 1959.

Shine, L. C., II & Bower, S. M. A one-way analysis of variance for single-subject designs. Educational and Psychological Measurement, 1971, 31, 105-113.

Siegel, S. Nonparametric statistics. New York: McGraw-Hill, 1956.

Strahan, R. F. A coefficient of directional correlation for time series analysis. Psychological Bulletin, 1971, 76, 211-214.

Tsao, Fei. General solution of the analysis of variance and co-variance in the case of unequal or disproportionate numbers of observations in subclasses. Psychometrika,

1946, 11, 107-128.

Vaughan, G. M. & Corballis, M. C. Beyond tests of significance: Estimating strength of effects in selected ANOVA designs. Psychological Bulletin, 1969, 72, 204-213.

Welch, B. L. The generalization of Student's problem when several different population variances are involved. Biometrika, 1947, 34, 28-35.

Wiener, N. Extrapolation, interpolation, and smoothing of stationary time series. New York: Wiley, 1949.

Winer, B. J. Statistical principles in experimental design. (2nd ed.) New York: McGraw-Hill, 1971.

Footnotes

[1]The author wishes to thank Donald P. Hartmann for his helpful comments on earlier drafts of this paper.

Portions of this paper were presented as a topical discussion at the meeting of the Southeastern Psychological Association, Atlanta, Georgia, March 1975.

[2]Requests for reprints should be sent to Michael R. Rosmann, Department of Psychology, Gilmer Hall, University of Virginia, Charlottesville, Virginia 22901.

[3]Methods of estimating the magnitude of a relationship between an independent and dependent variable in various ANOVA designs are available in recent articles by Dodd and Schultz (1973) and Vaughan and Corballis (1969). However, their relatively simple calculations only provide estimates of the magnitude of an effect and they say nothing about the direction or shape of an effect.

TABLE 1

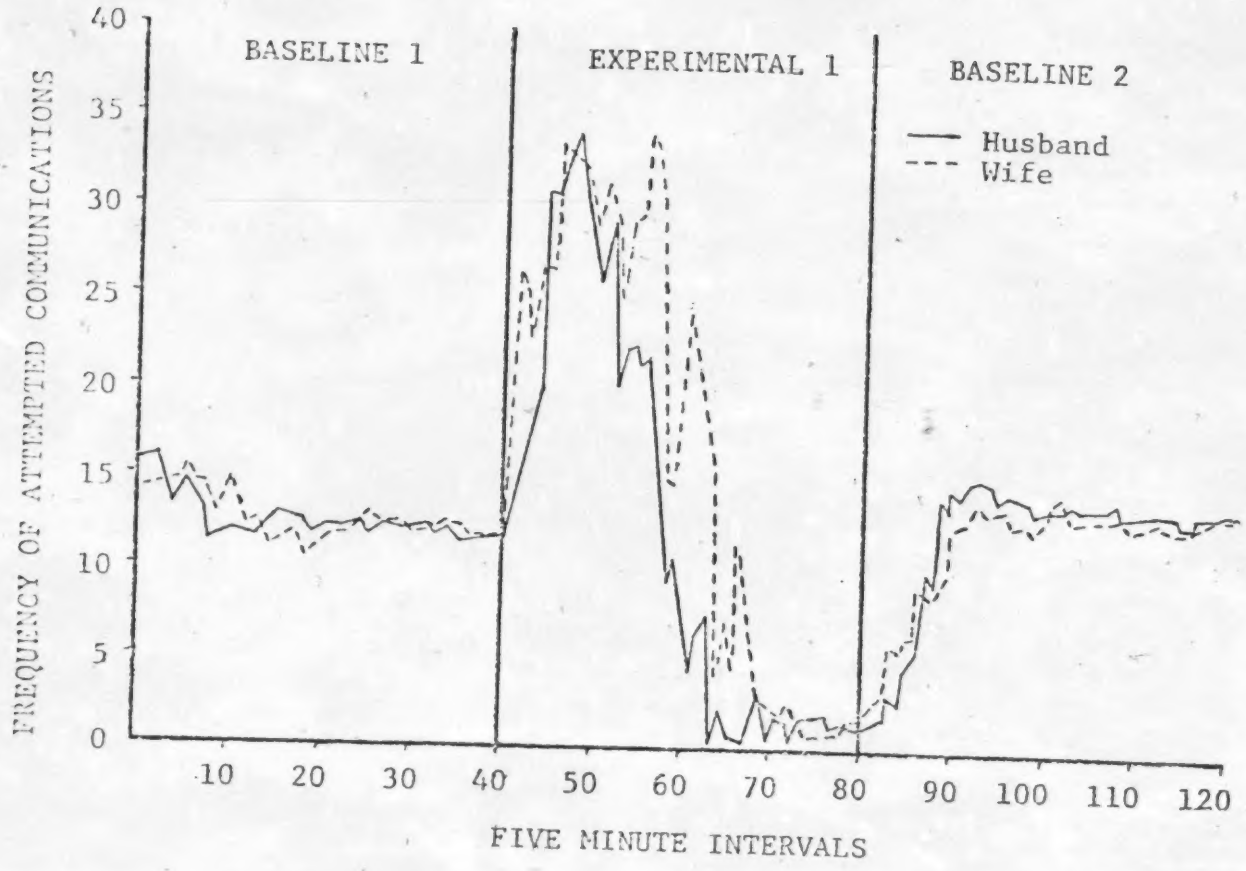
Frequency of Attempted Communications by a Husband and Wife to Each Other During a Hypothetical Experiment

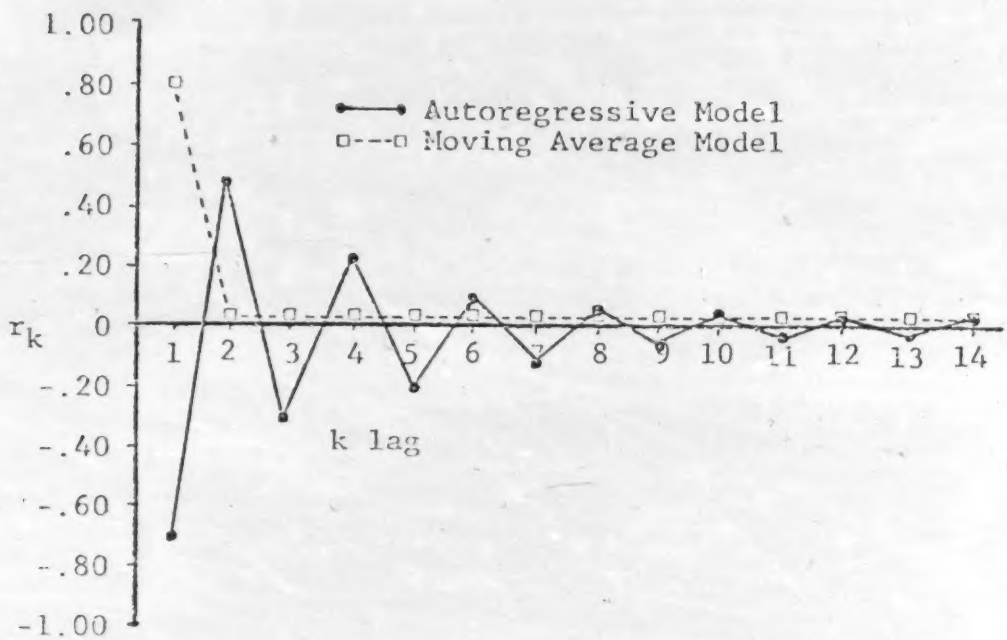
Subjects	Conditions		
	Baseline 1	Experimental 1	Baseline 2
Husband	n scores	n scores	n scores
Wife	n scores	n scores	n scores

Figure Captions

Figure 1. Hypothetical data illustrating the frequency of attempted verbal communications by a husband and wife to each other during a nondisrupted condition (Baseline 1 and 2) and a disrupted condition (Experimental 1).

Figure 2. Correlograms of a pure first order autoregressive model with $r_{lag = 1} = .7$ and a pure first order moving average model with $r_{lag = 1} = .8$.


BASELINE 1
EXPERIMENTAL 1
BASELINE 2
Husband
Wife
FREQUENCY OF ATTEMPTED COMMUNICATIONS
FIVE MINUTE INTERVALS
0
5
10
15
20
25
30
35
40
10
20
30
40
50
60
70
80
90
100
110
120


1.00
.80
.60
.40
.20
0
-.20
-.40
-.60
-.80
-1.00
r_k
1 2 3 4 5 6 7 8 9 10 11 12 13 14
k lag
Autoregressive Model
Moving Average Model